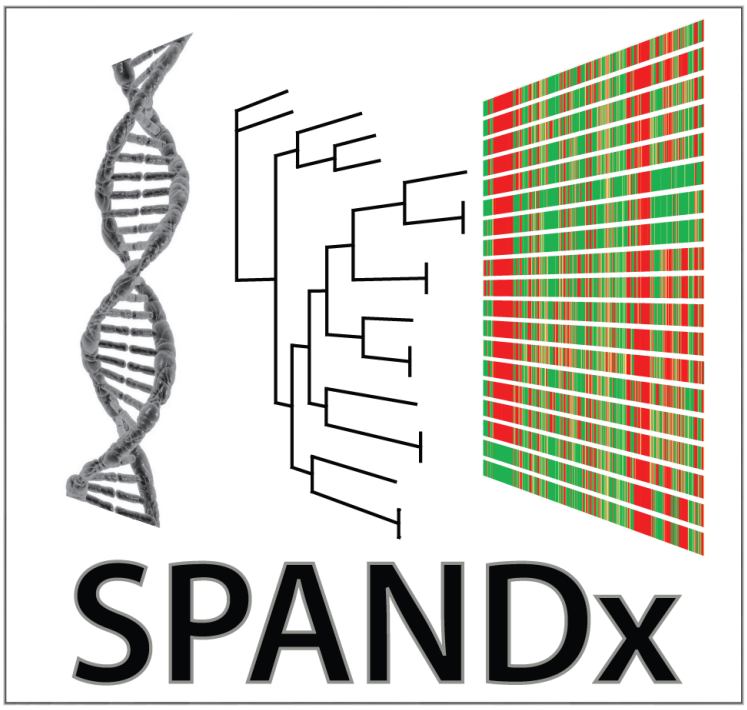


User Manual

Manual version 2.3, last modified 28Jul15

For the latest version of SPANDx and its associated user manual, please visit either of the following SPANDx websites: http://sourceforge.net/projects/spandx/ or https://github.com/dsarov/SPANDx

CONTENTS

Introduction and Description
Synopsis
Commands and Options
Installation and Requirements
Interpreting the Outputs
Examples
Authors and Citation
References

INTRODUCTION AND DESCRIPTION

SPANDx (**S**ynergised **P**ipeline for **A**nalysis of **N**ext-generation sequencing **D**ata in Linu**x**) is a comparative genomics pipeline designed to greatly simplify the identification of genetic variants (SNPs, insertions/deletions (indels), and large (>200bp) deletions) from medium- to large-sized haploid next-generation re-sequencing (NGS) datasets. SPANDx can currently process several NGS data inputs including paired- and single-end data from the Illumina MiSeq, HiSeq and GA_{IIx} platforms, single-end data from the Life Technologies Ion Personal Genome Machine (PGM)®, and single-end Roche 454 data. SPANDx integrates the following validated bioinformatics tools for start-to-finish sequence analysis of raw NGS data using a single command:

- **Burrows-Wheeler Aligner (BWA)** (1, 2) for alignment of short (i.e. Illumina and PGM) or long (i.e. 454) NGS read data. BWA is downloadable from: http://bio-bwa.sourceforge.net/. SPANDx does not currently support BWA versions later than 0.6.2.
- **SAMtools** (3) and **Picard** (as-yet-unpublished) for alignment manipulation and filtering. These programs can be downloaded from http://samtools.sourceforge.net/ and http://picard.sourceforge.net/, respectively.
- **Genome Analysis Tool Kit (GATK)** (4-6) for base quality score recalibration, realignment of regions with low mapping quality, duplicate removal, identification of SNPs and indels, and filtering the variant call format (VCF) file generated from the alignment process. GATK can be downloaded from http://www.broadinstitute.org/gatk/.
- **VCFtools** (7) for manipulation of VCF files. VCFtools is downloadable from: http://vcftools.sourceforge.net/. All file outputs from SPANDx are in standardised VCFv4.1 format. **tabix** and **bgzip** are VCFtools dependences required for handling `.vcf` files, and can be downloaded here: http://sourceforge.net/projects/samtools/files/tabix/.
- **BEDTools** (8), and specifically the *coverageBED* module, for identification of locus presence/absence across each genome of interest based on the reference sequence. This tool is useful for identifying larger-scale (approx. 100bp or larger) deletions. BEDTools can be downloaded from: https://github.com/arq5x/bedtools2/.
- **SnpEff** (9) for annotation of SNP and indel variants. SnpEff can be downloaded from here: http://snpeff.sourceforge.net/.
- **PLINK** (10) for microbial genome-wide association studies (mGWAS).

All of the above dependencies, with the exception of GATK and PLINK, come pre-bundled and pre-compiled with the SPANDx package. GATK needs to be downloaded and installed separately due to Broad Institute licencing restrictions. PLINK has not been included in the SPANDx bundle yet, but will be included in future versions.

The dependency binaries have been compiled for an x86-64 Linux system. If you have different system architecture, you will need to install SPANDx dependencies yourself.

Novel comparative genomic features of SPANDx include:

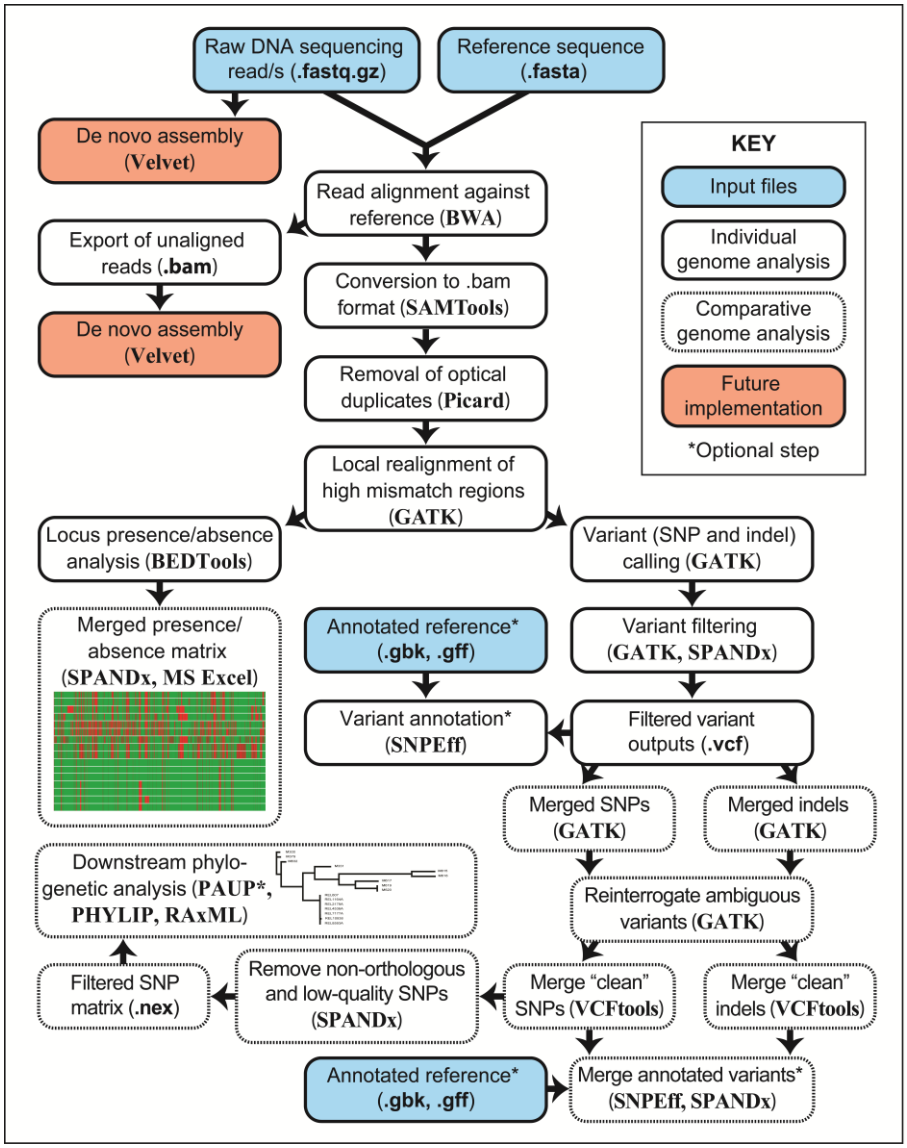
- Merged orthologous[1] SNP and indel matrices that greatly simplify variant visualisation for comparative genomic analyses.
- PAUP*, PHYLIP and RAxML-compatible `.nex` orthologous[1] SNP matrix files for downstream phylogenetic analyses. PHYLIP and RAxML are freely available programs downloadable from: http://evolution.genetics.washington.edu/phylip.html and http://www.exelixis-lab.org/. As of July 2015, PAUP* is now open source and will be frequently updated by the developer over the coming months. Please use Google to find the latest version!
- Locus presence/absence matrices from BEDTools outputs that enable simple visualisation and comparative genomic determination of 1) the 'core' genome and 2) variable genetic loci (including deleted regions brought about by reductive evolution).
- Merged, annotated SNP and indel matrices for fast and simple genetic characterisation of variants (NB. The user must provide SPANDx with the SnpEff-annotated reference genome information for variant annotation to work).

[1]The term "orthologous" refers to genetic loci that make up the 'core' genome. SNPs or indels residing in genetic loci that are missing in one or more genomes are excluded. If these variants are required, they can be found in the individual filtered `.vcf` output files.

Unlike many other comparative bioinformatics tools, SPANDx does not require pre-assembled genomes. In addition, the default settings for variant calling using Illumina, Ion PGM and 454 data have already been optimised and do not require the user to specify these settings, although these settings can be customised if required.

SPANDx has been purposely written for systems that utilise Portable Batch System (PBS), Sun Grid Engine (SGE) [i.e. qsub] or Simple Linux Utility for Resource Management (SLURM) [i.e. sbatch] to enable task parallelisation, greatly reducing turn-around-time of datasets comprising tens through to thousands of haploid genomes using a single command. As of version 2.7, SPANDx will also run on systems without a resource manager but will not run in parallel. Currently there is no support for SPANDx on other resource management systems (e.g. LSF) due to the unavailability of such systems in our laboratory, but compatibility with resource managers can be addressed if required. Please email us at mshr.bioinformatics@gmail.com if you require a specific resource manager compatible version of SPANDx and are willing to test it on your system.

The SPANDx workflow is shown below:


Raw DNA sequencing read/s (.fastq.gz)
Reference sequence (.fasta)
De novo assembly (Velvet)
Read alignment against reference (BWA)
Export of unaligned reads (.bam)
Conversion to .bam format (SAMTools)
De novo assembly (Velvet)
Removal of optical duplicates (Picard)
Local realignment of high mismatch regions (GATK)
KEY
Input files
Individual genome analysis
Comparative genome analysis
Future implementation
*Optional step
Locus presence/absence analysis (BEDTools)
Variant (SNP and indel) calling (GATK)
Merged presence/ absence matrix (SPANDx, MS Excel)
Annotated reference* (.gbk, .gff)
Variant filtering (GATK, SPANDx)
Variant annotation* (SNPEff)
Filtered variant outputs (.vcf)
Merged SNPs (GATK)
Merged indels (GATK)
Downstream phylo-genetic analysis (PAUP*, PHYLIP, RAxML)
Reinterrogate ambiguous variants (GATK)
Filtered SNP matrix (.nex)
Remove non-orthologous and low-quality SNPs (SPANDx)
Merge "clean" SNPs (VCFtools)
Merge "clean" indels (VCFtools)
Annotated reference* (.gbk, .gff)
Merge annotated variants* (SNPEff, SPANDx)

SYNOPSIS

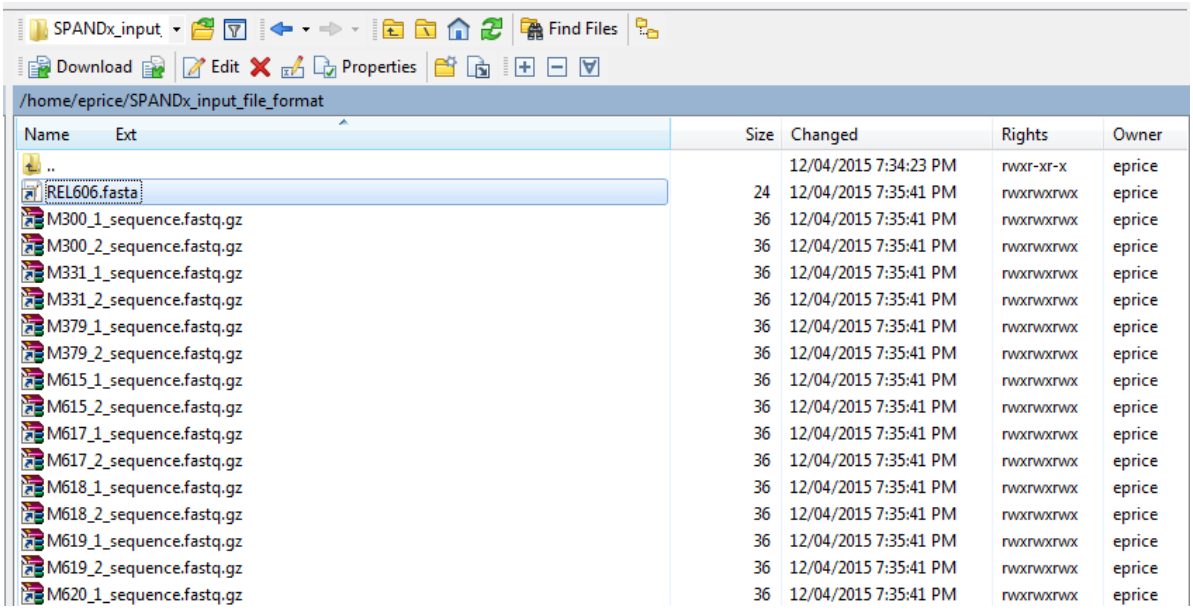
```
SPANDx -r <exact reference name, excluding .fasta extension> [parameters, optional] -o [organism] -m [generate SNP matrix yes/no] -i [generate indel matrix yes/no] -a [include annotation yes/no] -v [reference file for variant annotation; name must exactly match the SnpEff database name, which can be found in the snpEff.config file] -s [specify read prefix to run single strain; set to none to construct a SNP matrix from a previous analysis or leave as default to process all reads] -t [sequencing platform i.e. Illumina/Illumina_old/454/PGM] -p [pairing of reads (i.e. paired-end or single-end) PE/SE] -w [BEDTools window size in base pairs]
```

COMMANDS AND OPTIONS

`SPANDx.sh` is the only script that needs to be run to obtain data outputs. SPANDx by default expects paired-end Illumina data with v1.8+ quality encoding. If your data are in this format, the only required switch is `-r` to specify the reference sequence prefix. If another NGS data format is being analysed, please specify this format using the `-t` (and if single-end, the `-p`) switch/es[2]. By default, SPANDx *will* construct a locus presence/absence matrix but *will not* construct orthologous SNP or indel matrices, nor will it perform variant annotation. The `-m` (matrix) and `-i` (indel) switches are required for orthologous SNP and indel matrix creation, respectively. The `-a` (annotate) and `-v` (reference name for variant identification) switches are both required for variant annotation.

[2]SPANDx cannot process multiple NGS formats (e.g. single-end and paired-end Illumina) in a single run. If multiple NGS formats are to be analysed, please create separate analysis directories and run SPANDx specifically for each NGS format. These data can be merged for downstream analysis. See the `-s` description for more information.

Prior to running `SPANDx.sh`, both the reference (in `.fasta` format) and NGS files (in `.fastq.gz` format) are required to be in your analysis directory. SPANDx expects all NGS reads to be in the sequence analysis directory (i.e. the present working directory) and by default all NGS reads within the sequence analysis directory will be processed. Before running SPANDx, make sure the NGS read files conform to the following format regardless of the sequencing technology used: `strain_1_sequence.fastq.gz` and `strain_2_sequence.fastq.gz` (for paired-end reads) or `strain_1_sequence.fastq.gz` (for single-end reads). See screenshot below for correctly formatted paired-end input files.


SPANDx_input
Find Files
Download
Edit
Properties
/home/eprice/SPANDx_input_file_format
Name Ext
Size
Changed
Rights
Owner
..
12/04/2015 7:34:23 PM rwxr-xr-x eprice
REL606.fasta 24 12/04/2015 7:35:41 PM rwxrwxrwx eprice
M300_1_sequence.fastq.gz 36 12/04/2015 7:35:41 PM rwxrwxrwx eprice
M300_2_sequence.fastq.gz 36 12/04/2015 7:35:41 PM rwxrwxrwx eprice
M331_1_sequence.fastq.gz 36 12/04/2015 7:35:41 PM rwxrwxrwx eprice
M331_2_sequence.fastq.gz 36 12/04/2015 7:35:41 PM rwxrwxrwx eprice
M379_1_sequence.fastq.gz 36 12/04/2015 7:35:41 PM rwxrwxrwx eprice
M379_2_sequence.fastq.gz 36 12/04/2015 7:35:41 PM rwxrwxrwx eprice
M615_1_sequence.fastq.gz 36 12/04/2015 7:35:41 PM rwxrwxrwx eprice
M615_2_sequence.fastq.gz 36 12/04/2015 7:35:41 PM rwxrwxrwx eprice
M617_1_sequence.fastq.gz 36 12/04/2015 7:35:41 PM rwxrwxrwx eprice
M617_2_sequence.fastq.gz 36 12/04/2015 7:35:41 PM rwxrwxrwx eprice
M618_1_sequence.fastq.gz 36 12/04/2015 7:35:41 PM rwxrwxrwx eprice
M618_2_sequence.fastq.gz 36 12/04/2015 7:35:41 PM rwxrwxrwx eprice
M619_1_sequence.fastq.gz 36 12/04/2015 7:35:41 PM rwxrwxrwx eprice
M619_2_sequence.fastq.gz 36 12/04/2015 7:35:41 PM rwxrwxrwx eprice
M620_1_sequence.fastq.gz 36 12/04/2015 7:35:41 PM rwxrwxrwx eprice

Options:

`-r` *STR* *Required*. Specifies the reference genome file, excluding the `.fasta` extension. The `-r` switch is the only mandatory switch needed for SPANDx to function. Additional switches are required to modify the default behaviour of SPANDx and sequencing technology needs to be specified if your data are not paired-end Illumina data with v1.8+ quality scores. The reference file is required to be in `.fasta` format and should conform to the standard FASTA specification, as detailed here: http://www.ncbi.nlm.nih.gov/BLAST/blastcgihelp.shtml. IUPAC codes are not supported by some programs incorporated in SPANDx and *must* be avoided. For compatibility with the annotation module of SPANDx, the chromosome names for the reference genome must match those used by SnpEff. This nomenclature can be found in the `snpEff.config` file, which is generated upon SnpEff installation, or automatically with the full SPANDx installation. In addition, the `.fasta` reference file must not contain any blank lines.

`-o` *STR* *Optional.* Specifies the organism under analysis. The `-o` parameter is used in naming the read group headers in the SAM and BAM files. Spaces and special characters may have unexpected behaviour and should be avoided. [*Haploid*]

`-m` *yes/no* *Optional*. The `-m` switch is used to create a matrix with all orthologous SNP variants identified in the analysis. Non-orthologous SNPs are excluded. Output `.nex` files can be directly imported into PAUP*, PHYLIP or RAxML for phylogenetic analysis. By default this behaviour is switched off. [*no*]

`-i` *yes/no* *Optional*. The `-i` switch is used to create a matrix with all orthologous indel variants identified in the analysis. Non-orthologous indels are excluded. By default this behaviour is switched off. [*no*]

`-a` *yes/no* *Optional*. The `-a` switch is used to perform annotation of the variant files. By default this behaviour is switched off. If annotation is switched on the `-v` switch must also be specified. [*no*]

`-v` *STR* *Optional, required if* `-a` *is set to "yes".* The `-v` switch is used to specify the reference genome that SnpEff will use to annotate variants. The string used for this variable must match one of the genomes contained within the `SnpEff.config` file. Additionally, chromosome names in the reference file must match those contained within the SnpEff program. Please refer to the SnpEff manual (which can be found here: http://snpeff.sourceforge.net/SnpEff_manual.html) for more information. [*null*]

`-s` *STR* *Optional*. The `-s` switch is used to flag a single genome for analysis. If `-s` is set to `none`, SPANDx will not perform individual analysis of any NGS read files in the current directory. Instead, SPANDx will move to the comparative genomics section of the pipeline (see SPANDx workflow above) and assume all individual genome analysis has already been completed. SPANDx will then merge all VCF files contained in `$PWD/Phylo/snps` and perform error correction using the `.bam` and `.bai` files contained in `$PWD/Phylo/bams` to construct an orthologous SNP matrix, which will be output to `$PWD/Phylo/out`. Before running this module, please check that all `.vcf` files located within `$PWD/Phylo/snps` match the alignment files within `$PWD/Phylo/bams`, and

that all `.bam` files contain their accompanying `.bai` index file. This feature mitigates the need to re-run previous analyses from scratch, and is useful for combining data generated from multiple SPANDx runs (e.g. from different sequencing technologies) into a single orthologous SNP matrix. [*all*]

`-t` *STR* *Optional.* The `-t` switch specifies the sequence technology used, and must be one of the following: `Illumina`, `Illumina_old`, `454` or `PGM`. By default, SPANDx expects Illumina reads with Phred+33 read quality encoding, which is standard as of v1.8+. To specify NGS reads generated by the older Illumina format (i.e. Phred+64), use `-t Illumina_old`. If the analysis mode is switched to `454` or `PGM`, SPANDx will use the BWA-SW algorithm of BWA for read alignment and thus will expect reads to be in single-end `strain_1_sequence.fastq.gz` format with Phred+33 quality encoding. [*Illumina*]

`-p` *STR* *Optional*. The `-p` switch specifies the pairing of reads and must be either `PE` or `SE`. By default SPANDx expects reads to be paired. If reads are single end, `-p` *must* be set to `SE`. Currently SPANDx does not support paired-end 454 or PGM data. [*PE*]

`-w` *INT* Optional. The `-w` switch specifies the window size (in base pairs) used by BEDTools to analyse whole genome alignment coverage i.e. locus presence/absence. [*1000*]

INSTALLATION AND REQUIREMENTS

SPANDx is written in Bash and will run on most Linux installations. For parallelisation, SPANDx can utilise PBS, SGE or SLURM resource managers. SPANDx has been tested on GNU Bash version 3.2.25(1)-release and GNU Bash version 4.1.2(1)-release (both on x86_64-redhat-linux-gnu) with Java v1.7.0_55 and v1.7.0_71. The 2.7 version of SPANDx (downloadable here: http://sourceforge.net/projects/spandx/ or here: https://github.com/dsarov/SPANDx) has been tested using PBS (TORQUE v2.5.13), SGE v6.2u5p3, and SLURM v14.11. Dependency versions tested in SPANDx v2.7 are BWA v0.6.2, SAMtools v0.1.19, Picard v1.105, the Genome Analysis Toolkit v3.0, BEDTools v2.18.2, SnpEff v4.1, VCFtools v0.1.11 and tabix v0.2.6. SPANDx does not currently support BWA versions later than 0.6.2.

> For parallelisation, SPANDx requires PBS, SGE or SLURM to submit jobs to the cluster. If you do not have one of these system setups, SPANDx cannot be run in parallel. Please contact us if you are using a different Linux resource scheduler setup and are willing to trial-and-error SPANDx on your system.

SPANDx should work from any installation directory but has been most extensively tested in `/home/user/bin`. To install SPANDx, gunzip and untar the program (usually with the command `tar xvfz SPANDx_full.tar.gz`) in your `/bin` directory. Optionally, SPANDx now has its own Github page, meaning it can be installed directly using the following command: `git clone https://github.com/dsarov/SPANDx`

IMPORTANT: following extraction of the script files, the `SPANDx.config` file will need to be modified to contain the location of the SPANDx installation. If your system uses a proxy to access the internet, please modify the JAVA_PROXY variable in `SPANDx.config`. For default SPANDx behaviour, no other settings should need to be modified. In addition, you will also need to download and install GATK and either place the GenomeAnalysisTK.jar file (renamed without version numbers) in the SPANDx installation directory or specify the install path in the `SPANDx.config` file.

For users who don't have x86-64 systems: You will need to download and install SPANDx dependencies yourself as only x86-64 binaries are included in the SPANDx distribution. Following extraction of the script files, the `SPANDx.config` file will need to be modified to contain the absolute paths of each dependency (i.e. BWA, SAMTools, Picard, GATK, BEDTools, SnpEff, VCFtools and tabix installations – see **Introduction and Description** above for web links to these free third-party programs). If a "dependency not found" error occurs, please check the installation path and the location specified in the `SPANDx.config` file.

Please make sure to specify the location of the PERL5 libraries (automatically installed with VCFtools) in the `SPANDx.config` file. PERL5 libraries are required for correct functioning of VCFTools. Make sure that the location of tabix and bgzip (dependencies of VCFtools) are specified in your `PATH` variable.

Some of these programs have additional dependencies that are required for them to function properly (e.g. Java). Please refer to the appropriate manuals or system administrator for installation of these utilities if they are not already on your Linux system.

SPANDx customisation**:**

Resource manager:

Depending on your cluster environment, the `scheduler.config` file may need to be changed. By default, SPANDx will expect the resource manager to be PBS. **IMPORTANT: If you are using SGE, please modify the `SCHEDULER` variable to `SGE`. If you are using SLURM, please modify the `SCHEDULER` variable to `SLURM`. If no resource manager is available, please modify the `SCHEDULER` variable to `NONE`.** This file will configure the operation of the resource manager. By default, all commands request one node and 12 hours of wall time, you will not be sent mail, and standard output is merged with standard error. These settings can be changed if your job needs more time to complete or if you want e-mail notifications of job completion.

Variant calling:

One advantage of SPANDx over other tools is that GATK variant calling parameters are already specified. These parameters have been tested across NGS data for several bacterial species generated on different NGS platforms. Therefore, the default settings should work well for most haploid genome projects. If desired, users can customise the SPANDx variant filtering parameters by altering the `GATK.config` file. All filtering steps used in GATK can be customised using this configuration file. Note that these variables must conform to JEXL specifications.

> SPANDx variant calling has been optimised for bacterial genomes and therefore may behave differently for other haploid organisms. If in doubt, outputs should be verified with e.g. wet lab analysis of variants. SPANDx is currently not configured to analyse diploid or polyploid genomes.

The following parameters can be customised to change the variant filtering behaviour for both SNPs and indels if required; below are the default SPANDx settings:

`CLUSTER_SNP=3` (for SNPs only)
`CLUSTER_WINDOW_SNP=10` (for SNPs only)
`MLEAF=0.95`
`QD=10.0`
`MQ=30.0`
`FS=10.0`
`HAPLO=20.0`
`QUAL=30.0`
`LOW_DEPTH=2` (variants with less than the average coverage, divided by `LOW_DEPTH`, will fail filtering. If this value is set at the default of 2, regions with less than half the average depth will fail and thus will be filtered out).
`HIGH_DEPTH=3` (a value of 3 means that regions with more than three times the average coverage of the entire genome will fail and thus will be filtered out).

INTERPRETING THE OUTPUTS

Unaligned reads:

A `.bam` file of the unaligned reads is generated by SAMTools after BWA alignment. The unaligned reads can be found in: `$PWD/strain/unique/unmapped.bam`. It is anticipated that future versions of SPANDx will include an option for automated assembly of these unaligned reads.

Alignment files:

Alignment files generated with SPANDx are in `.bam` format and are found in: `$PWD/strain/unique/strain.realigned.bam`. If visualisation of the alignment is desired, these files can be easily viewed in an alignment viewer (our favourite is Tablet (11); downloadable from: http://ics.hutton.ac.uk/tablet/).

Whole genome coverage (a.k.a. locus presence/absence):

Following assessment of whole genome coverage by BEDTools, SPANDx provides a combined BEDcov matrix for all analysed genomes in: `$PWD/Outputs/Comparative/Bedcov_merge.txt`. This file lists the BEDTools 'windows', or NGS read coverage, for each strain based on the reference sequence, and ranges from 0 (0% read coverage across the window) to 1 (100% coverage across the window). The BEDcov matrix can be imported into Microsoft Excel for easier visualisation and manipulation. `Bedcov_merge.txt` is a useful file for identifying the 'core' genome of a given dataset and for identifying variable genetic regions among strains. The bp resolution of this output can be modified by changing the `-w` switch (e.g. 100 *cf*. the default 1000). We suggest changing to 100 for microbial GWAS analysis in PLINK to increase sensitivity, and leaving as default for all other analyses.

Variants:

SNPs and indels are output from SPANDx analysis into two locations: 1) `$PWD/Outputs/SNPs_indels_PASS`, which contains both SNP and indel variants that have passed the filters specified in the `GATK.config` file (see "**Variant calling**" above for details of the default filters); and 2) `$PWD/Outputs/SNPs_indels_FAIL,` which contains SNPs and indels that have failed filtering parameters. If annotation is switched on, annotated variants will be output to: `$PWD/strain/unique/annotated`. Annotated SNP and indel outputs will be generated for each genome under analysis. In addition, if both annotation and comparative analysis is switched on (`-a` yes and `-m` yes), annotated, merged SNP and indel matrices are generated for all genomes under analysis. These files are found in `$PWD/Outputs/Comparative` and are called `All_indels_annotated.txt` and `All_SNPs_annotated.txt`. These are tab-delimited text files that can be easily imported into Excel, as shown in the screenshot below. Note that the binary column will need to be specified as "Text only" due to the character string containing "0". The binary column is a representation of the SNP/indel pattern at that specific location, which can be useful for filtering.

Location	K96243	MSHR5662	MSHR5667	MSHR5670	Binary_code	Effect	Impact	Functional_Class	Codon_change	Amino_Acid_change	Gene_name
1_1534	G	A	A	A	0222	NON_SYNONYMOUS_CODING	MODERATE	MISSENSE	aCg/aTg	T281M	BPSL0002
1_14099	C	T	T	T	0222	SYNONYMOUS_CODING	LOW	SILENT	ctC/ctT	L30	gspG
1_16320	A	G	G	G	0222	UPSTREAM	MODIFIER	-	-	71	gspK
1_16382	G	C	C	C	0222	UPSTREAM	MODIFIER	-	-	9	gspK
1_16764	T	C	C	C	0222	NON_SYNONYMOUS_CODING	MODERATE	MISSENSE	tTg/tCg	L125S	gspK
1_16789	A	G	G	G	0222	SYNONYMOUS_CODING	LOW	SILENT	caA/caG	Q133	gspK
1_17077	T	C	C	C	0222	SYNONYMOUS_CODING	LOW	SILENT	gtT/gtC	V229	gspK
1_17636	G	A	A	A	0222	NON_SYNONYMOUS_CODING	MODERATE	MISSENSE	Gcg/Acg	A51T	gspL
1_65151	C	T	T	.	022.	SYNONYMOUS_CODING	LOW	SILENT	gaG/gaA	E199	BPSL0059
1_65664	A	G	G	G	0222	SYNONYMOUS_CODING	LOW	SILENT	atT/atC	I28	BPSL0059
1_340085	C	T	T	T	0222	STOP_GAINED	HIGH	NONSENSE	tgG/tgA	W16*	BPSL0319
1_340115	C	T	T	T	0222	SYNONYMOUS_CODING	LOW	SILENT	gcG/gcA	A6	BPSL0319
1_340333	T	C	C	C	0222	SYNONYMOUS_CODING	LOW	SILENT	gcA/gcG	A235	BPSL0320
1_341120	T	C	C	C	0222	NON_SYNONYMOUS_CODING	MODERATE	MISSENSE	Agc/Ggc	S458G	BPSL0321

N.B. SNPs/indels represented with "?" are an ambiguous call and should be interpreted with caution. SNPs/indels represented with "." do not pass the depth filters and are likely in deleted regions.

SPANDx can repeat the variant filtering steps without repeating the alignment and data processing steps. To use this behaviour, remove the relevant `snps.PASS` and `indels.PASS` files from the `$PWD/Outputs/SNPs_indels_PASS` directory and the relevant `snps.AMB` and `indel.AMB` files from the `$PWD/Outputs/SNPs_indels_FAIL` directory, change the `GATK.config` file to the desired parameters and re-run SPANDx. NB. SPANDx will only re-filter the variants with altered parameters for those strain/s that have been removed from the `Output` directories.

Orthologous SNP matrices for phylogenetic analyses:

Two separate SNP matrix files are generated by SPANDx. These matrices are output in `$PWD/Outputs/Comparative/` and are named `Ortho_SNP_matrix_RAxML.nex` and `Ortho_SNP_matrix.nex`.

> The `Ortho_SNP_matrix` files generated by SPANDx exclude SNPs that are low-quality, that are in non-orthologous regions, and that are tri- or tetra-allelic in nature. Non-orthologous SNPs should not be used for phylogenetic reconstruction. In addition, filtering for tri- and tetra-allelic SNPs is performed to minimise erroneous calls (which can look like tri- and tetra-allelic SNPs) passing through filters.
>
> NB. The annotated SNP and indel matrices (`All_indels_annotated.txt` and `All_SNPs_annotated.txt`) include tri-allelic variants. In addition, ambiguous and non-orthologous calls are flagged as "?" and ".", respectively.

`Ortho_SNP_matrix.nex` includes SNP coordinates identified by GATK. This file is directly importable into PAUP* and is useful for phylogenetic estimations that require nucleotide data (e.g. maximum likelihood). Below is a screen capture of this file:

```
#nexus
begin data;
dimensions ntax=17 nchar=106568;
format symbols="AGCT" gap=. datatype=nucleotide transpose;
taxlabels REL606 M300 M331 M379 M615 M617 M618 M619 M620 M632 REL10938 REL1164A REL2179A REL4536A REL607 REL7177A REL8593A;
matrix
NC_012967_58 G G G G G G G G C G G G G G G
NC_012967_64 T T T T T T T T C T T T T T T
NC_012967_67 C C C C C C T C C C C C C C C
NC_012967_171 T A T A T T T T T T T T T T T
NC_012967_392 G T G T G G G T T G G G G G G
NC_012967_437 C C C C C C G C C C C C C C C
NC_012967_464 C C C C C C T C C C C C C C C
NC_012967_473 C C C C T C T C C C C C C C C
NC_012967_479 G G G G A G A G G G G G G G G
NC_012967_509 G G G G G G A G G G G G G G G
NC_012967_536 C C T C C C C C C C C C C C C
NC_012967_558 T T C T C T C T T T T T T T T
NC_012967_587 A G A G A A A G G A A A A A A
NC_012967_590 C C C C T C T C C C C C C C C
NC_012967_620 T T C T C C C T T C T T T T T
NC_012967_632 C C T C C C C C C C C C C C C
NC_012967_659 T T T T C T T T T T T T T T T
NC_012967_668 G G A G G G G G G A G G G G G
NC_012967_689 C C C C C C C C C T C C C C C
```

Ortho_SNP_matrix_RAxML.nex is a RAxML- and PHYLIP-importable version of the Ortho_SNP_matrix_RAxML.nex file. Note that for compatibility with PHYLIP, taxa names must have exactly 10 characters (including spaces) otherwise Ortho_SNP_matrix_RAxML.nex will not be recognised as a valid PHYLIP file. SPANDx does not automatically rename taxa to meet this PHYLIP requirement.

Merged SNP-indel matrices:

In certain instances, particularly in outbreak studies where only closely related strains are being examined and there are few genetic variants, it may be desirable to merge SNP and indel variants prior to phylogenetic analysis for increased resolution. We have recently shown that this approach can provide more robust phylogenetic correlation with epidemiological data than constructing a phylogeny based on SNPs alone (12). To merge SNP and indel variants into a separate output file, run the MergeSnpIndel.sh script while in the $PWD/Outputs/Comparative/ directory. The MergeSnpIndel.sh script will merge and reformat the data in the Ortho_SNP_matrix.nex and indel_matrix.nex files. The output file is: indel_SNP_matrix.nex. This file is directly importable into PAUP*. Below is a screen capture of this file:

```
#nexus
begin data;
dimensions ntax=17 nchar=107451;
format symbols="01" gap=. datatype=standard transpose;
taxlabels REL606 M300 M331 M379 M615 M617 M618 M619 M620 M632 REL10938 REL1164A REL2179A REL4536A REL607 REL7177A REL8593A;
matrix
NC_012967_58 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0
NC_012967_64 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0
NC_012967_67 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0
NC_012967_171 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0
NC_012967_392 0 1 0 1 0 0 0 1 1 0 0 0 0 0 0 0 0
NC_012967_437 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0
NC_012967_464 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0
NC_012967_473 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0
NC_012967_479 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0
NC_012967_509 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0
NC_012967_536 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
NC_012967_558 0 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0 0
NC_012967_587 0 1 0 1 0 0 0 1 1 0 0 0 0 0 0 0 0
NC_012967_590 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0
NC_012967_620 0 0 1 0 1 1 1 0 0 1 0 0 0 0 0 0 0
NC_012967_632 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
NC_012967_659 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
NC_012967_668 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0
```

Microbial genome-wide association studies (mGWAS):

The main comparative outputs of SPANDx (SNPs, indels and presence/absence matrices in `$PWD/Outputs/Comparative/`) can be used as input files for mGWAS. As of version 2.6, SPANDx is distributed with `GeneratePlink.sh`.

The `GeneratePlink.sh` script requires two input files: an `ingroup.txt` file and an `outgroup.txt` file. The `ingroup.txt` file should contain a list of the strains of interest (e.g. antibiotic-resistant strains) and the `outgroup.txt` file should contain a list of all strains lacking the genotype or phenotype of interest (e.g. antibiotic-sensitive strains). Although larger taxon numbers in the ingroup and outgroup files will increase the statistical power of mGWAS, it is better to only include relevant strains i.e. do not include strains that have not yet been characterised. The `ingroup.txt` and `outgroup.txt` files ***must*** include only one strain per line and ***must*** be in UTF-8 text file format; do not save in other formats. This script will generate `.ped` and `.map` files for SNPs, and presence/absence loci and for indels if these were identified in the initial analyses. The `.ped` and `.map` files can be directly imported into PLINK. For more information on GWAS and how to run PLINK, please refer to the PLINK website: http://pngu.mgh.harvard.edu/~purcell/plink/

Log files:

By default, both the standard error and standard output are merged into a single log file. Almost all commands in SPANDx are captured in log files. If an error occurs, the log files are a good first place to look.

If you wish to minimise the amount of log files that SPANDx generates, you can change the `ERROR_OUTPUT` variable (PBS) to `n`, or the `ERROR_OUTPUT_SGE` (SGE) to `no`, in the `scheduler.config` file. Please note that this feature only works when using PBS or SGE resource handlers.

EXAMPLES

The simplest way to run SPANDx is if your reads are in paired-end, Illumina format and follow the naming convention of `strain_1_sequence.fastq.gz` and `strain_2_sequence.fastq.gz`. SPANDx can then be run by simply specifying the reference `.fasta` genome prefix. All read files in the current directory will be processed, although a SNP or indel matrix will not be constructed nor will variant annotation be performed unless specified.

"No frills" SPANDx command for basic Illumina 1.8+ analysis:

```
/home/user/bin/SPANDx/SPANDx.sh -r reference
```

If other SPANDx features are required or reads other than Illumina v1.8+ are used, these features will need to be specified as per the examples below.

Paired-end Illumina 1.8+ reads, SNP matrix required, no annotated genome available/required:

```
/home/user/bin/SPANDx/SPANDx.sh -r reference -m yes
```

To include an annotation for the above example:

```
/home/user/bin/SPANDx/SPANDx.sh -r reference -a yes -m yes -v
ref_genome_in_SnpEff_database
```

Paired-end Illumina 1.8+ reads, indel matrix required, no annotated genome available/required:

```
/home/user/bin/SPANDx/SPANDx.sh -r reference -i yes -m yes
```

Paired-end Illumina 1.3 reads, SNP and indel matrices required, annotated reference genome Hi_86-028NP.fasta available:

```
/home/user/bin/SPANDx/SPANDx.sh -r Hi_86-028NP -o Hi -m yes -i yes -a
yes -v Haemophilus_influenzae_86_028NP_uid58093 -t Illumina_old
```

Single-end Ion PGM reads, SNP/indel matrices and annotation not required, BEDCov window size of 500bp (instead of the default 1000bp) desired, using the same reference genome as above:

```
/home/user/bin/SPANDx/SPANDx.sh -r Hi_86-028NP -o Hi -t PGM -p SE -w 500
```

Paired-end Illumina 1.3 reads, annotated genome available/required for the reference genome Hi_86-028NP.fasta with a single strain (Hi_00345) for alignment and variant calling. No SNP/indel matrices required:

```
/home/user/bin/SPANDx/SPANDx.sh -r Hi_86-028NP -o Hi -a yes -v
Haemophilus_influenzae_86_028NP_uid58093 -t Illumina_old -s Hi_00345
```

mGWAS (generation of input data compatible with PLINK)

Running PLINK for mGWAS analysis across antibiotic-resistant strains vs. antibiotic-sensitive strains:

```
/home/user/bin/SPANDx/GeneratePlink.sh -i
inGroup_antibiotic_resistant.txt -o OutGroup_antibiotic_sensitive.txt
```

Running PLINK for mGWAS analysis as above but changing locus presence-absence cutoffs for customised outputs (default is 0.9):

```
/home/user/bin/SPANDx/GeneratePlink.sh -i
inGroup_antibiotic_resistant.txt -o OutGroup_antibiotic_sensitive.txt -c
0.95
```

AUTHORS AND CITATION

SPANDx was developed by Dr. Derek Sarovich and Dr. Erin Price from Menzies School of Health Research, Darwin, NT 0810, Australia.

Derek's homepage: http://www.menzies.edu.au/page/Our_People/Researchers/Derek_Sarovich/

Erin's homepage: http://www.menzies.edu.au/page/Our_People/Researchers/Erin_Price/

If you find an error or bug, please contact Derek or Erin at mshr.bioinformatics@gmail.com. Please include a detailed description of the error you encountered, the operating system you used and what happened to cause the error. In addition, please send the appropriate log files with the description.

Derek can be followed on Twitter @DerekSarovich (https://twitter.com/dereksarovich). Any feedback regarding SPANDx is most welcome!

If you used SPANDx and found it useful for your research, please cite it! ☺

Sarovich DS and Price EP. 2014. SPANDx: a genomics pipeline for comparative analysis of large haploid whole genome re-sequencing datasets. *BMC Res. Notes* **7**:618.

REFERENCES

1. **Li H, Durbin R.** 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25:**1754-1760.
2. **Li H, Durbin R.** 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics **26:**589-595.
3. **Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup GPDP.** 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics **25:**2078-2079.
4. **McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA.** 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research **20:**1297-1303.
5. **DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ.** 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genetics **43:**491-498.
6. **Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, Gabriel S, DePristo MA.** 2013. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics **11:**11 10 11-11 10 33.
7. **Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, Genomes Project Analysis G.** 2011. The variant call format and VCFtools. Bioinformatics **27:**2156-2158.
8. **Quinlan AR, Hall IM.** 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics **26:**841-842.
9. **Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM.** 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly (Austin) **6:**80-92.
10. **Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC.** 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet **81:**559-575.
11. **Milne I, Stephen G, Bayer M, Cock PJ, Pritchard L, Cardle L, Shaw PD, Marshall D.** 2013. Using Tablet for visual exploration of second-generation sequencing data. Briefings in Bioinformatics **14:**193-202.
12. **McRobb E, Sarovich DS, Price EP, Kaestli M, Mayo M, Keim P, Currie BJ.** 2015. Tracing melioidosis back to the source: using whole-genome sequencing to investigate an outbreak originating from a contaminated domestic water supply. J Clin Microbiol **53:**1144-1148.